Clinical use of the Hamilton Depression Rating Scale: is increased efficiency possible? A post hoc comparison of Hamilton Depression Rating Scale, Maier and Bech subscales, Clinical Global Impression, and Symptom Checklist-90 scores

Ruhe, H.G.; Dekker, J.J.M.; Peen, J.; de Jonghe, F.

published in
Comprehensive Psychiatry
2005

DOI (link to publisher)
10.1016/j.comppsych.2005.03.001

document version
Publisher's PDF, also known as Version of record

Available online at www.sciencedirect.com

SCIENCE DIRECT

Comprehensive Psychiatry 46 (2005) 417–427

Comprehensive PSYCHIATRY

www.elsevier.com/locate/comppsych

Clinical use of the Hamilton Depression Rating Scale: is increased efficiency possible? A post hoc comparison of Hamilton Depression Rating Scale, Maier and Bech subscales, Clinical Global Impression, and Symptom Checklist-90 scores☆

Henricus G. Ruhé[a,b,*], Jack J. Dekker[a,c], Jaap Peen[a], Rebecca Holman[d], Frans de Jonghe[a,b]

[a]*Mentrum, Depression Research Group, PO Box 75848, 1070 AV Amsterdam, The Netherlands*
[b]*Department of Psychiatry, Academic Medical Center, PO Box 22660, 1100 DD Amsterdam, The Netherlands*
[c]*Department of Clinical Psychology, Vrije Universiteit Amsterdam, 1081 BT Amsterdam, The Netherlands*
[d]*Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, PO Box 22660, 1100 DD Amsterdam, The Netherlands*

Abstract

Background: The 17-item Hamilton Depression Rating Scale (HDRS) is used as a semi–gold standard in research. In treatment guidelines, the HDRS measurements serve to determine response and remission and guide clinical decision making for nonresponders. However, its use in clinical practice is limited, possibly because the HDRS is time consuming. In addition, the multidimensional HDRS is criticized for not measuring a unidimensional aspect as depression severity. The Maier and the Bech, two 6-item severity subscales extracted from the HDRS, are relatively unknown. This paper investigates whether the measurements obtained with these subscales are comparable with the original HDRS measurements.

Methods: Data from 2 randomized controlled trials in 482 male and female patients, diagnosed with a major depression (with or without dysthymia) according to *Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition*, of whom 219 participated in the trials, were reanalyzed. A standardized stepwise psychopharmacological treatment was compared with a combination of pharmacotherapy with Short Psychodynamic Supportive Psychotherapy in a psychiatric outpatient department. Outcome measures were internal consistency and concurrent validity of HDRS, Maier, Bech, Clinical Global Impression scales, and Symptom Checklist depression subscale. Effect sizes of HDRS, Maier, and Bech were used to compare measured treatment effects for the randomized subjects participating in the trials. Item Response Theory was used to obtain conversion tables for the HDRS, Maier, Bech, and Symptom Checklist depression subscale.

Results: We found moderate internal consistency (Cronbach $\alpha \approx 0.6$-0.7) and high correlations of the Maier and Bech subscales with overall HDRS scores. Overall, there were no clinically relevant differences in effect sizes between Maier, Bech, and HDRS, although some differences were statistically significant. Receiver operating characteristic curves showed no difference between Maier and Bech to define remission but showed the Clinical Global Impression ratings to be unreliable. A cutoff ≤ 4 corresponded with an HDRS ≤ 7 criterion in both subscales.

Conclusion: In clinical practice, both Maier and Bech scales can be used as equivalents of the HDRS, but will be more efficient.

1. Introduction

Major depressive disorder is a severe disabling illness, expected to be the world's second health problem in 2020 [1]. Depression is associated with high costs, regarding direct treatment and indirect costs of loss of productivity and quality of life [2]. Several clinical guidelines were developed to guide the treatment of this disorder; both psychotherapy and pharmacotherapy (or in combination) appear effective [3-10].

The use of self-report or clinician-rated symptom scales is recommended to assess severity and response to treatment [8,11,12]. Some experts claim clinician-rated symptom scales to have a larger validity and reliability than self-reporting scales, especially in patients with cognitive impairment, and more severe or psychotic depressions

☆ Conflicts of interest. External funding did not support these post hoc analyses.

* Corresponding author. Department of Mood Disorders, MFO Psychiatrie AMC/De Meren, 1105 BC Amsterdam, The Netherlands. Tel.: +31 20 5662240; fax: +31 20 6919139.

E-mail address: h.g.ruhe@amc.uva.nl (H.G. Ruhé).

0010-440X/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.comppsych.2005.03.001

[11,13,14]. Specific symptom scales are more reliable than global rating scales [11,13,15] Especially, rating scales can be used to objectively determine specific cutoff points for response and remission [12,16,17].

In most clinical trials, the Hamilton Depression Rating Scale (HDRS) [18,19]—a clinician-rated symptom scale—is used as a standard to determine severity and response. [5,8,11,15,20-23]. Many versions of the HDRS exist, with the number of items usually varying between 17 and 24 [11,18,19,22]; however, up to 36 items have been described [23]. Longer versions were especially developed to cover reverse neurovegetative (atypical) symptoms [23]. The Clinical Global Impression (CGI) [24]—a clinician-rated global scale—is also frequently used [5,8,15,25]. In clinical practice, although recommended, rating scales are not used routinely. Explanations for this discrepancy could be ignorance of existing scales, a strong belief in one's clinical judgement, an unsystematic approach of depression, and also the amount of time needed for rating scales (eg, 15-20 minutes for the HDRS [11]) and the necessity of training [20,26].

The HDRS is criticized as being sensitive to somatic symptoms (eg, somatic illness or side effects of drugs) [11,15,27,28], for not rating all 9 *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* domains, its unequal weightings of different symptoms, and for the multidimensionality of the HDRS total score [13,21,29-31]. Multidimensionality is important to cover the maximum range of clinical features of major depressive disorder but does not necessarily measure depression severity. Multidimensional scales can be misleading when measurement of severity and treatment response is concerned [13,21,28], especially when the measured depressive symptoms do not change proportionally with depression severity. Finally, some reports emphasize that the HDRS systematically favors (sedative) tricyclic antidepressants (TCAs) above selective serotonin reuptake inhibitors (SSRIs) [27,32-35]. Sleep and somatic items may appear to be "improved" by side effects of TCAs but worsened by side effects (eg, insomnia, gastrointestinal complaints, and agitation) of SSRIs.

To overcome the problems of the multidimensional HDRS mentioned above, a more unidimensional subscale from the HDRS covering core symptoms of severity is desired. Also, from a clinical point of view, fewer items will be less time consuming for application by busy clinicians. However, for the purpose of reference, subscale scores must remain anchored to the original HDRS. To identify shorter unidimensional subscales, Maier and Philipp [28] used Rasch and Mokken analyses, and Gibbons et al [29] used factor analysis. Bech et al [36] developed another 6-item subscale. This scale initially emerged from an analysis with experienced psychiatrists as a validity criterion [36] and was validated psychometrically thereafter using Rasch analyses [37,38]. This Bech subscale was combined with 4 items of the Cronholm-Ottosson Depression Scale to form the Bech-Rafaelsen Melancholia Scale [39]. Santor and Coyne [21] examined the score performances of individual HDRS items as a function of depression severity with a nonparametric Item Response Theory (IRT) approach, retaining 14 items. These 14 items included all 6 items of the Maier subscale and all 8 items of the Gibbons subscale. However, 1 item from the Bech subscale (13, somatic symptoms) was not included.

In a meta-analysis of individual patient data, Faries et al [40] evaluated the responsiveness of total HDRS and subscale scores in TCA and SSRI pharmacotherapy trials, finding a maximal sensitivity for the Maier subscale. In a similar reanalysis, Entsuah et al [41] found larger effect sizes (E-S) for the Bech, Maier, and Gibbons subscales compared with the HDRS in trials comparing SSRIs or venlafaxine. O'Sullivan et al [20] found comparable sensitivity to detect changes for the 6-item Bech subscale compared with the 17-item HDRS. Hooper and Bakish [26] found equal sensitivity to change during treatment for the 6-item Bech subscale compared with the HDRS 17-item version. Moller [32] and Bech et al [42-44] used the Bech subscale to reexamine treatment efficacy of SSRIs and mirtazapine (vs TCAs or placebo). The latter publications did not provide data for the Maier subscale.

In this paper, we describe a secondary analysis of our trial data to answer the following questions:

(1) Are the Maier, Bech, and HDRS comparable in the measurement of depression severity and the sensitivity to measure changes in severity?
(2) Is this comparability stable across the full range of response to treatment (eg, nonresponse, partial response, and full response), across different treatments and different baseline severity of depression?
(3) What are clinical cutoff points for the subscales to determine remission compared with conventional definitions [12,16,17].

We hypothesized that the differences between Maier, Bech, and HDRS scales would be small and that there would be no apparent effect modification across neither treatments nor baseline severity. In contrast, we hypothesized that for nonresponse and partial responders, the E-S would be smaller than for responders. This would additionally prove the hypothesis of sensitivity to change.

2. Method

2.1. Patient selection

In the present analyses, we use data from 2 published, randomized controlled trials conducted between 1993 and 1998 [45,46]. The first trial aimed at efficacy and effectiveness of pharmacotherapy versus the combination of pharmacotherapy with Short Psychodynamic Supportive Psychotherapy (SPSP) [47-50] (16 sessions) [45]. The second trial investigated efficacy and effectiveness of a combination of pharmacotherapy with 8 versus 16 sessions of SPSP [46]. Pharmacotherapy in both trials consisted of

3 successive steps in case of intolerance or inefficacy. Both trials started with fluoxetine (20 mg/d), when this was unsuccessful (Clinical Global Impression Improvement [CGI-I] >2, only "minimally improved" or worse) after 6 weeks, amitriptyline (≥150 mg/d, dependent of plasma levels) was initiated in trial 1 and nortriptyline (≥150 mg/d, dependent of plasma levels) in trial 2. If again unsuccessful after 6 weeks, moclobemide (300-600 mg/d) was started in trial 1 and mirtazapine (30-45 mg/d) in trial 2.

Inclusion criteria for participation in the trials were age between 18 and 60 years, *Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition*-defined major depression (with or without dysthymia) assessed in a structured clinical interview, a 17-item HDRS baseline score of at least 14 points, and written informed consent. Patients were excluded in case of psycho-organic or psychotic or dissociative disorders, drug abuse, or when the patient was considered to be too unreliable to participate in a clinical trial. Other Axis 1 comorbidity was not excluded. Further exclusion criteria were if there was a serious communicative or practical problem (eg, language barrier or the patient will soon leave the country), if there was a contraindication for 1 of the antidepressants used, if the patient was adequately treated with antidepressants during the present depressive episode, if the patient used other psychotropic medication, or if the patient was or planned to become pregnant. Additional exclusion criteria were of the usual kind in drug research: "too ill" (eg, antidepressants must be started immediately) and/or "too suicidal" (eg, hospitalization is unavoidable) to participate in a clinical trial. The study was approved by the medical ethics committee. After complete description of the study to the subjects, written informed consent was obtained.

Of 3226 newly registered outpatients, 988 patients had a depressive disorder. By initial screening, 503 of these 988 patients were excluded by the above exclusion criteria leaving 485 subjects (including patients that later refused to participate or had an HDRS below 14; further referred to as the *diagnostic sample*). To enter the trials, a second exclusion check was performed by a psychiatrist (excluding 73 patients), and 142 subjects with an HDRS-17 <14 were excluded, leaving 270 patients for randomization. After randomization, 51 patients refused participation, leaving 219 patients who started the proposed therapy (further referred to as the *per protocol sample*) [45,46].

In this manuscript, we used the diagnostic sample for most cross-sectional analyses, and the randomized patients in the per protocol sample for analyses of sensitivity of response data. For noncompleters, the last observation was carried forward (LOCF).

2.2. Outcome measures

Primary outcome measures were the 17-item HDRS [18,19], the Maier subscale of the HDRS (containing items 1, 2, and 7-10) [28], the Bech subscale of the HDRS (items 1, 2, 7, 8, 10, and 13) [37], the Clinical Global Impression Severity (CGI-S) and Improvement (CGI-I) scale [24], and the Symptom Checklist depression subscale ($SCL\text{-}90_{dep}$) [51,52]. Thus, 3 levels of information were obtained: data from (1) an independent, trained, supervised, and blinded research assistant (HDRS-17, Maier, and Bech), (2) the treating clinician (CGI-S/I), and (3) the patient ($SCL\text{-}90_{dep}$). The HDRS was administered using a semistructured interview [53]. Before participating in the study, the reliability of the HDRS assessments was established. During the study, to avoid slippage, audiotaped assessments were discussed monthly.

In the analyses of treatment efficacy, response was defined as a ≥50% HDRS score reduction, partial response as ≥20% to 50% reduction in HDRS score, and remission as an HDRS score of ≤7 points [16,54].

2.3. Statistics

Cronbach α coefficients and mean inter-item correlations were used to express internal consistency. To check whether the increased number of items in the HDRS accounted for a higher Cronbach α coefficient than in the subscales (with only 6 items), we applied the Spearman-Brown formula [55]. Next, we calculated concurrent validity as Pearson correlation coefficients between total HDRS, Maier, and Bech subscale scores and $SCL\text{-}90_{dep}$ scores. Linear regression models calculated variance of HDRS scores explained by the subscales [56]. These analyses were performed in our diagnostic sample. Concurrent validity between CGI-S/I and HDRS subscale ratings was determined also, however, to avoid low correlations because of limited dispersion; this was done for the last observation in the per protocol sample. The CGI improvement scale was compared with changes expressed as percentages of the baseline score.

To compare differences in sensitivity to measure treatment effects (also referred to as responsiveness) in data from the per protocol sample, E-S for HDRS, Maier, and Bech subscales were calculated per subject as the within-subject changes in scale scores divided by the pooled standard deviation of the mean change in scale score

$$\left(\frac{T_0 - T_{\text{end(LOCF)}}}{SD_{\text{pooled-difference}}}\right)$$

[20]. In this way, differences in E-S could be tested and 95% confidence intervals (95% CIs) could be calculated. Differences in E-S between the scales were tested by paired *t* tests. To determine significant effect modification, the above analyses were repeated while data were stratified. For stratification, we used initial HDRS scores of at least 19 for severe depression [11], criteria for response as described above, and treatment condition. Differences in E-S between strata were tested by analyses of variance (ANOVAs) models.

The Partial Credit IRT model [57] was used to estimate the relationships between total scores on the HDRS and total

scores on the Maier and Bech subscales of the HDRS. The scores were those obtained at exit (per protocol sample). The computer program OPLM (One Parameter Logistic Model; CITO-group, Arnhem, The Netherlands) [58] was used to obtain a set of weights for each item in the HDRS using conditional maximum likelihood methods. The same software and the item weights were used to obtain estimates of the latent trait associated with each score on the HDRS, the Maier subscale, and the Bech subscale. The total scores for the pairs of scales were equated by matching the total scores for which the latent trait scores were most similar [59]. These methods are very similar to those used in a previous publication about the Quick Inventory of Depressive Symptomatology (QIDS) [60]. The range of SCL scores associated with each HDRS score was obtained directly from the original data.

Finally, receiver operating characteristic (ROC) curves were constructed to summarize validity of cutoff points.

Table 1
Studied populations

	Diagnostic sample (n = 485[a])	Per protocol sample (n = 219)[b] Trial I [45]		Trial II [46]		Combined I + II
		AD (n = 57)	AD + SPSP (16) (n = 72)	AD + SPSP (8) (n = 45)	AD + SPSP (16) (n = 45)	
Sex (% female)	60.3	63.2	61.1	60.0	68.9	63.0
Age	35.3 ± 9.9	34.9 ± 8.2	34 ± 9.4	38.1 ± 10.5	36.2 ± 10.5	35.5 ± 9.7
Marital status (%)						
Married	97 (20.2)	12 (21.1)	10 (13.9)	19 (42.2)	9 (20.5)	50 (22.9)
Divorced	60 (12.5)	7 (12.3)	8 (11.1)	5 (11.1)	9 (20.5)	29 (13.3)
Widowed	3 (0.6)	–	–	1 (2.2)	–	1 (0.5)
Never married	318 (66.1)	38 (66.7)	54 (75.0)	20 (44.4)	26 (59.1)	138 (63.3)
Other	3 (0.6)	–	–	–	–	–
Educational level (%)						
Low	72 (15)	11 (19.3)	13 (18.3)	8 (17.8)	8 (18.2)	40 (18.4)
Intermediate	179 (37.3)	21 (36.8)	22 (31.0)	20 (44.4)	16 (36.4)	79 (36.4)
High	229 (47.7)	25 (43.9)	36 (50.7)	17 (37.8)	20 (45.5)	98 (45.2)
Occupational (%)						
Job	166 (34.7)	18 (31.6)	22 (30.6)	19 (42.2)	14 (31.8)	73 (33.5)
On sickness	134 (28.0)	14 (24.6)	23 (31.9)	13 (28.9)	16 (36.4)	66 (30.3)
Social security	84 (17.5)	12 (21.1)	10 (13.9)	8 (17.8)	3 (6.8)	33 (15.1)
Disabled	27 (5.6)	3 (5.3)	5 (6.9)	1 (2.2)	3 (6.8)	12 (5.5)
Student	41 (8.6)	5 (8.8)	10 (13.9)	2 (4.4)	3 (6.8)	20 (9.2)
Other	27 (5.6)	5 (8.8)	2 (2.8)	2 (4.4)	5 (11.4)	14 (6.4)
Duration of episode (y)						
<1	314 (67.0)	39 (70.9)	49 (70.0)	27 (61.4)	25 (56.8)	140 (65.7)
1-2	70 (14.9)	6 (10.9)	9 (12.9)	7 (15.9)	14 (31.8)	36 (16.9)
>2	85 (18.1)	10 (18.2)	12 (17.1)	10 (22.7)	5 (11.4)	37 (17.4)
Psychiatric treatment during this episode (%)	95 (20.2)	13 (23.6)	9 (12.9)	5 (11.4)	10 (22.7)	37 (17.4)
Antidepressants last 3 mo before study (%)	77 (16.4)	6 (10.9)	10 (14.3)	9 (20.5)	9 (20.5)	34 (16.0)
Of which adequate (%)	1 (0.2)	–	–	–	–	–
Depressive episodes (previous 5 y)						
None	296 (63.1)	26 (47.3)	4 (58.6)	33 (75.0)	28 (65.1)	128 (60.4)
1-2	138 (29.4)	21 (38.2)	22 (31.4)	10 (22.7)	13 (30.2)	66 (31.1)
3 or more	35 (7.5)	8 (14.5)	7 (10.0)	1 (2.3)	2 (4.7)	18 (8.5)
Ethnicity (%)						
Northwest European	414 (86.3)	51 (89.5)	63 (87.5)	38 (84.4)	40 (90.9)	192 (88.1)
Mediterranean	18 (3.8)	3 (5.3)	2 (2.8)	1 (2.2)	1 (2.3)	7 (3.2)
Caribbean	22 (4.6)	1 (1.8)	2 (2.8)	5 (11.1)	2 (4.5)	10 (4.6)
Other	26 (5.4)	2 (3.5)	5 (6.9)	1 (2.2)	1 (2.3)	9 (4.1)
Baseline scores of rating scales						
HDRS-17	17.1 ± 6.5	21.0 ± 4.8	20 ± 4.9	19.4 ± 3.8	20.3 ± 4.4	20.2 ± 4.6[c]
Maier	9.2 ± 3.6	11.0 ± 2.9	10.9 ± 2.8	10.7 ± 2.3	10.9 ± 2.9	10.9 ± 2.8[c]
Bech	9.4 ± 3.7	11.5 ± 2.8	11.2 ± 2.7	10.7 ± 2.3	11.0 ± 3.0	11.1 ± 2.3[c]
CGI-S[d]	4.7 ± 0.7	4.8 ± 0.6	4.7 ± 0.7	4.5 ± 0.7	4.6 ± 0.6	4.7 ± 0.7
SCL-90$_{dep}$ subscale	45.9 ± 11.8	48.7 ± 11.7	47.8 ± 9.8	49.3 ± 8.7	52.0 ± 10.1	49.2 ± 10.2[c]

Data represent means (±SD) unless indicated. Denominators of percentages vary because of missing values.

[a] Total diagnostic sample.

[b] No significant differences between treatment groups (per protocol sample) (ANOVA or χ^2).

[c] Significant differences ($P < .05$) between included and excluded patients (independent t test).

[d] n = 241.

Table 2
Internal validity and concurrent validity of HDRS-17, Maier, Bech, SCL-90 depression subscale, and CGI-S

	Internal consistency		Concurrent validity: Pearson r (% explained variance)					
	Cronbach α	Mean inter-item correlation	Overall		Moderate depression[a]		Severe depression[a]	
			Maier	Bech	Maier	Bech	Maier	Bech
Diagnostic sample								
Maier[b]	0.62	0.21	–	–	–	–	–	–
Bech[b]	0.67	0.25	0.95 (91%)	–	0.91 (83%)	–	0.90 (81%)	–
HDRS-17[b]	0.73	0.13	0.86 (75%)	0.86 (73%)	0.57 (57%)	0.76 (58%)	0.65 (42%)	0.60 (36%)
SCL-90 Depression subscale[b]	0.88	0.33	0.64 (41%)	0.64 (40%)	0.45 (21%)	0.45 (20%)	0.40 (16%)	0.42 (18%)
Per protocol sample								
CGI-S end point[c]	NA		0.54 (29%)	0.58 (34%)	0.55 (31%)	0.55 (30%)	0.59 (34%)	0.65 (42%)
CGI-I end point[d]	NA		0.42 (18%)	0.43 (18%)	0.38 (14%)	0.37 (13%)	0.48 (23%)	0.49 (24%)

[a] Severe depression defined as initial HDRS-17 $\geq$19 (n = 221).
[b] Maier, Bech, and HDRS: n = 482; SCL-90$_{dep}$: n = 473.
[c] CGI-S: n = 229.
[d] Compared with change expressed as percentage of baseline rating.

Differences in areas under the curve (AUC) were tested with attention for interrelation (because we studied these tests within the same subjects) as described by Hanley and McNeil [61]. For all data analyses except the IRT analysis, SPSS for Windows version 10.1 was used [62]. For all tests, 2-tailed significance levels were applied.

Table 3
Pretreatment and posttreatment Maier, Bech, and HDRS scores with corresponding E-S in per protocol sample

	Mean ± SD baseline	Mean ± SD end point (LOCF)	Mean decrease (95% CI)	SD of decrease	E-S (95% CI)
All subjects (n = 219)					
Maier	10.9 ± 2.75	6.2 ± 4.46	4.7 (4.1-5.3)	4.54	1.03 (0.89-1.16)[a,b]
Bech	11.1 ± 2.69	6.2 ± 4.50	4.9 (4.3-5.5)	4.54	1.08 (0.95-1.22)
HDRS-17	20.2 ± 4.56	12.0 ± 7.62	8.2 (7.2-9.2)	7.45	1.10 (0.96-1.23)
Moderate depression (initial HDRS-17 <19; n = 93)[c]					
Maier	8.9 ± 2.13	5.3 ± 3.86	3.7 (2.8-4.5)	4.19	0.81 (0.62-1.00)[b]
Bech	9.2 ± 2.12	5.1 ± 3.89	4.1 (3.2-5.0)	4.28	0.91 (0.71-1.10)
HDRS-17	16.2 ± 1.42	9.8 ± 6.15	6.4 (5.1-7.8)	6.47	0.86 (0.68-1.04)
Severe depression (initial HDRS-17 $\geq$19; n = 126)[c]					
Maier	12.3 ± 2.25	6.9 ± 4.75	5.4 (4.6-6.2)	4.66	1.19 (1.01-1.37)[a]
Bech	12.5 ± 2.16	7.0 ± 4.75	5.5 (4.7-6.3)	4.65	1.21 (1.03-1.39)
HDRS-17	23.1 ± 3.79	13.7 ± 8.18	9.5 (8.1-10.8)	7.88	1.27 (1.08-1.46)
Final nonresponders (n = 65)[d]					
Maier	10.6 ± 2.82	10.6 ± 2.86	0.0 (−0.7-0.6)	2.65	−0.01 (−0.15-0.14)[e]
Bech	11.1 ± 2.76	10.7 ± 2.89	0.4 (−0.2-1.1)	2.66	0.09 (−0.05-0.24)[e]
HDRS-17	19.5 ± 4.25	19.9 ± 4.55	−0.4 (−1.2-0.3)	3.08	−0.06 (−0.16-0.05)[e]
Final partial responders (n = 64)[d]					
Maier	11.5 ± 2.86	7.4 ± 3.07	4.1 (3.4-2.7)	2.58	0.90 (0.76-1.04)
Bech	11.4 ± 2.59	7.4 ± 3.19	4.1 (3.4-4.7)	2.71	0.90 (0.75-1.04)
HDRS-17	21.4 ± 5.06	14.2 ± 4.41	7.2 (6.7-7.7)	1.88	0.96 (0.90-1.03)
Final responders (n = 90)[d]					
Maier	10.6 ± 2.59	2.1 ± 2.00	8.5 (7.8-9.1)	3.12	1.87 (1.72-2.01)[f]
Bech	10.9 ± 2.71	2.1 ± 1.89	8.8 (8.1-9.4)	3.14	1.93 (1.79-2.08)
HDRS-17	19.9 ± 4.28	4.8 ± 3.46	15.1 (14.1-16.1)	4.89	2.02 (1.89-2.16)[f]

Stratification by depression severity and final treatment response. Note that the overlap of two 95% CI of E-S does not rule out a statistical significant difference between these E-S (see text).
[a] Significantly different from E-S HDRS (paired t test; $P < .05$).
[b] Significantly different from E-S Bech (paired t test; $P < .05$).
[c] Significant differences of E-S Maier, E-S Bech, and E-S HDRS between moderate and severe depression (ANOVA; $P < .05$).
[d] Response criteria: decrease in HDRS scores: <20% = nonresponse, 20%-50% partial response, and $\geq$50% = response. Significance differences of E-S Maier, E-S Bech, and E-S HDRS between categories of response (ANOVA; $P < .001$).
[e] Significant differences between E-S Maier–E-S Bech, E-S HDRS–E-S Maier, and E-S HDRS–E-S Bech (paired t test; $P < .05$).
[f] Significant difference between E-S HDRS–E-S Maier (paired t test; $P < .05$).

3. Results

3.1. Patient characteristics

Table 1 shows demographics for the diagnostic and per protocol samples. There were no significant differences observed between the diagnostic and per protocol sample (tested as excluded vs included), except from a lower mean HDRS score (and Maier, Bech, and SCL-90 depression scores) in the diagnostic sample. This difference was caused by the application of the entrance criterion (HDRS $\geq$14) for randomization. No significant differences existed between the different treatment groups. The studied population existed of mainly unmarried, mid-30s, moderately to highly educated, female, white adults, with moderate to severe depressive episodes of less than 1-year duration. More than 75% of the subjects were not treated for the current depressive episode before; 16% received an inadequate trial of an antidepressant.

3.2. Internal and concurrent validity

Data for internal and concurrent validity are presented in Table 2. Cronbach α's were slightly lower for the Maier and Bech subscales. If a 17-item scale is reduced to 6 items, the expected α is 0.49 (Spearman-Brown formula). Thus, the observed values of 0.62 and 0.60 show increased internal validity for the subscales. The mean inter-item correlation was markedly higher for the Maier and Bech subscales. The correlation between Maier and Bech subscales was high. Both Maier and Bech subscales explained approximately 75% of the variance of the total HDRS score. The self-rated SCL-90_{dep} was reasonably well correlated with the HDRS (r = 0.67) and the Maier and Bech subscales (r = 0.64). Concurrent validity of the scales was overall slightly less in the more depressed subgroup (HDRS $\geq$19; n = 194) compared with moderately depressed subjects, except for the correlation between HDRS and Maier subscale. The CGI-S at study end point was moderately correlated with the HDRS (r = 0.57), as with the Maier and Bech subscales. The CGI-S showed higher correlation with the Bech subscale, especially in those severely depressed. The CGI-I at study end point was less well correlated with the percentage change in HDRS (r = 0.42) and the subscales.

3.3. Sensitivity to change

In Tables 3 and 4, overall and stratified E-S in the per protocol sample are presented. In these tables, the 95% CI of the E-S indicates whether the E-S significantly deviates from 0 (no effect measured). Comparisons between E-S may produce significant differences between E-S, even when the 95% CIs between the 2 E-S overlap. Of the 9 comparisons between the Maier and Bech subscales made in these tables, 5 were not significant. The Maier was significantly different from the HDRS in 4 of 9 comparisons, whereas the Bech was significantly different from the HDRS in only 1 of the 9 comparisons. Differences between E-S were small.

In the total per protocol sample, the Maier subscale was significantly less powerful to observe treatment effects: the E-S assessed by the Maier was significantly lower than the E-S of the Bech and HDRS. When stratified for depression severity, the E-S of Maier, Bech subscales, and HDRS were larger in severe compared with moderate depression. A significant difference between these strata was observed for all E-S (ANOVA). Within the group of severely depressed

Table 4
Pretreatment and posttreatment Maier, Bech, and HDRS scores with corresponding E-S in per protocol sample

	Mean ± SD baseline	Mean ± SD end point (LOCF)	Mean decrease (95% CI)	SD of decrease	E-S (95% CI)
All subjects (n = 219)					
Maier	10.9 ± 2.75	6.2 ± 4.46	4.7 (4.1-5.3)	4.54	1.03 (0.89-1.16)[a,b]
Bech	11.1 ± 2.69	6.2 ± 4.50	4.9 (4.3-5.5)	4.54	1.08 (0.95-1.22)
HDRS-17	20.2 ± 4.56	12.0 ± 7.62	8.2 (7.2-9.2)	7.45	1.10 (0.96-1.23)
AD (n = 57)[c]					
Maier	11.0 ± 2.95	7.2 ± 4.97	3.8 (2.4-5.1)	5.12	0.83 (0.53-1.13)[d]
Bech	11.5 ± 2.78	7.1 ± 5.01	4.4 (3.1-5.7)	5.01	0.97 (0.68-1.26)[d]
HDRS-17	21.0 ± 4.77	13.9 ± 8.36	7.1 (4.9-9.3)	8.38	0.95 (0.65-1.25)
AD + 8 SPSP (n = 45)[c]					
Maier	10.7 ± 2.34	5.9 ± 4.31	4.8 (3.6-5.9)	3.81	1.05 (0.80-1.31)
Bech	10.7 ± 2.28	6.0 ± 4.21	4.6 (3.5-5.8)	3.87	1.02 (0.76-1.28)
HDRS-17	19.4 ± 3.80	11.1 ± 6.80	8.3 (6.4-10.2)	6.37	1.12 (0.86-1.38)
AD + 16 SPSP (n = 117)[c]					
Maier	10.9 ± 2.82	5.8 ± 4.21	5.1 (4.2-5.9)	4.47	1.12 (0.93-1.30)
Bech	11.1 ± 2.78	5.8 ± 4.32	5.3 (4.4-6.1)	4.54	1.16 (0.98-1.35)
HDRS-17	20.1 ± 4.69	11.5 ± 7.44	8.6 (7.3-10.0)	7.37	1.16 (0.98-1.34)

Stratification by treatment modality. Note that the overlap of two 95% CI of E-S does not rule out a statistical significant difference between these E-S (see text). AD indicates antidepressants.

[a] Significantly different from E-S HDRS (paired t test; $P < .05$).

[b] Significantly different from E-S Bech (paired t test; $P < .05$).

[c] No significant differences of E-S Maier, E-S Bech, and E-S HDRS between treatment modalities (ANOVA).

[d] Significant differences between E-S Maier–E-S Bech (paired t test; $P < .05$).

Table 5
The conversion between the HDRS total scores and the Maier subscale, Bech subscale, and the range of SCL scores using IRT analysis (per protocol sample)

HDRS	Maier	Bech	Range SCL-90$_{dep}$
0	0	0	–
1-2	1	1	–
3-4	2	2	–
5	3	3	–
6	3	4	–
7 (Remission[a])	4	4	–
8-9	5	5	–
10-11	6	6	–
12	7	7	–
13 (Mild[a])	7	8	–
14	8	8	31-61
15	8	9	22-59
16	9	9	26-61
17	9	10	25-59
18 (Moderate[a])	10	10	30-62
19	10	10	28-60
20	11	11	38-67
21	11	11	30-72
22	12	12	39-61
23	12	12	36-61
24 (Severe[a])	13	13	38-67
25 (Very severe[a])	13	13	45-64
26	13	13	47-71
27	14	14	46-72
28	14	14	42-79
29	15	14	55-71
30	15	15	43-43
31	15	15	63-70
32	16	15	62-65
33	16	16	57-71
34-35	17	16	–
36	17	17	–
37-39	18	17	–
40-42	19	18	–
43-44	20	19	–
45	21	19	–
46	21	20	–
47-48	22	20	–

The only valid conversions that can be made from this table are between (1) HDRS and Maier, (2) HDRS and Bech, and (3) HDRS and SCL-90$_{dep}$.
[a] Cutoffs as provided by Yonkers and Samson [11].

subjects, the Maier was significantly less sensitive compared with the HDRS (paired t test). Within the moderately depressed group, the Bech outperformed the Maier (paired t test). Across different strata of final response, significant differences in E-S were found (ANOVA). Within strata, the Bech subscale performed less in final nonresponders, whereas the Maier performed significantly less than the total HDRS in final responders (paired t tests).

In Table 4, it is shown that no overall differences in E-S were found between treatment modalities (ANOVA). Within the group of patients treated with antidepressants only, the Maier subscale was significantly less sensitive to detect treatment differences than the HDRS; however, the Maier did not differ significantly from the Bech subscale (paired t test).

3.4. Conversion of HDRS scores, criteria for remission, and depression severity

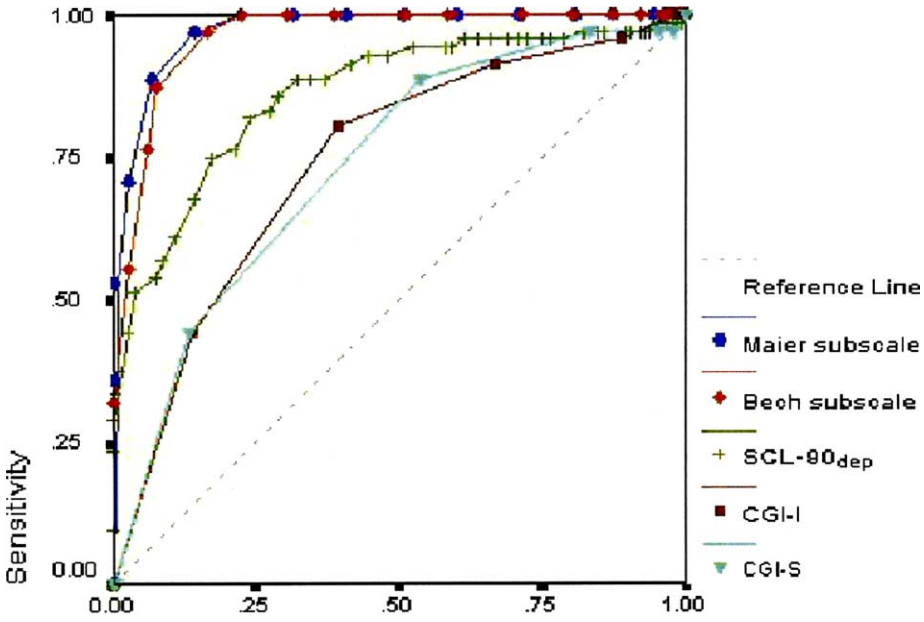
Table 5 shows the conversion between HDRS scores and Maier, Bech, and SCL-90$_{dep}$ scores. Maier and Bech cutoff scores to define remission, mild, moderate, and severe depression [11]) can be identified. Fig. 1 shows the ROC curves for Maier, Bech CGI-S, CGI-I, and SCL-90$_{dep}$ cutoff scores, with HDRS ≤ 7 as the reference criterion. The difference in AUC for the Maier and Bech subscales was not significant ($z = 1.25$; $P = .21$). The difference in AUC between Maier and Bech subscales compared with SCL-90$_{dep}$ and both CGIs was highly significant ($z > 3.8$; $P < .001$). In the table below Fig. 1 sensitivity and specificity for cutoff scores ≤ 3 and ≤ 4 for the Maier and Bech subscales are given.

4. Discussion

4.1. Major findings

This study examined the relative effectiveness of the HDRS subscales as developed by Maier and Philipp [28] and Bech et al [37] in monitoring severity and treatment effects in depression. We found that the Maier and Bech subscales gave results comparable to the original 17-item HDRS, with high concurrent validity and increased mean inter-item correlations and internal consistency. Maier and Bech subscales were highly comparable to each other in the measurement of treatment changes. Differences between E-S were rather small and clinically irrelevant. For interpretation, a conversion table linking HDRS scores and Maier and Bech scores were provided. The Maier had a slightly (nonsignificant) higher sensitivity and specificity to predict the reference criterion for remission (HDRS ≤ 7). Both Maier and Bech subscales differentiated nonresponders from partial and final responders.

A significant difference in sensitivity to change existed between the Bech and Maier within the group treated with antidepressants only. We were unable to find the reason for this difference compared with other treatment modalities, where the difference between Maier and Bech was not found or was not significant. The question arises whether there is a difference in sensitivity between the Maier and Bech subscales across different treatment modalities or that other (postrandomization) differences between the groups or mere chance explains this observation. Because this difference was not found in the other groups (treated with both antidepressants and psychotherapy), we think it cannot be ascribed to a difference in detecting pharmacological (side) effects. If a Bonferroni correction would be applied for the number



	Maier		Bech	
Cutoff	Sens.	Spec.	Sens.	Spec.
Subscale =3	88.9	93.2	87.5	92.5
Subscale =4	97.2	85.7	97.2	83.7

Fig. 1. ROC curves for Maier and Bech subscales, SCL-90 depression, and CGI-S/I at end point compared with HDRS-17 "Remission" in per protocol sample. HDRS-17 score ≤ 7 (remission). Sens indicates sensitivity; Spec, specificity AUC (SE): Maier 0.972 (0.009), Bech 0.963 (0.011), SCL-90_{dep} 0.862 (0.028), CGI-S 0.743 (0.036), and CGI-I 0.738 (0.035).

of comparisons tested ($P < .01$), the observed difference would not maintain its significance.

4.2. The relevance of the difference between the Maier and Bech subscales

The only difference between the Maier and Bech subscales is the inclusion of agitation (eg, running thoughts or restlessness, 0-4 points) in the Maier versus the inclusion of general somatic symptoms (eg, tiredness, 0-2 points) in the Bech. It could be argued that one scale is comparable with the other scale without the different item; for example, the Maier subscale would then be comparable to the Bech subscale minus the "general somatic" item. In our (diagnostic) sample, the item agitation contributed 1.3 (SD = 0.9) points to the total Maier score (9.2, SD = 3.6). The general somatic item contributed 1.5 (SD = 0.8) points to the total Bech score (9.4, SD= 3.7). Thus, overall tiredness was more present than agitation in this sample, and agitation was not rated near its maximum like tiredness. Both items occurred intraindividually at the same time but were not interchangeable. This means that the Maier and Bech subscales show different perspectives on depressive symptoms. In this respect, it is noteworthy to mention that the agitation item was dropped beforehand when the Bech subscale was developed and validated, because this item showed limited variance (ie, was not found to be scored) in the 2 studied samples [36,37]. Furthermore, in the Maier subscale, the items psychomotor agitation and psychomotor retardation are included, which—at first sight—seem to represent 2 opposed polarities. However, these items also co-occurred within the same individuals. This can be explained by the broad definitions of agitation (both restlessness or running thoughts) and retardation (both retardation in activities or in thinking) in our semistructured interview.

The original HDRS is often criticized to measure somatic symptoms [11,15,27,28]. Although the Bech subscale was designed as an unidimensional scale, the "general somatic" item is still among the 6 items. However, in the Rasch analysis, this item was the least contributive and showed a ceiling effect for moderate and severe depression [37]. Although both a *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* and *International Statistical Classification of Diseases, 10th revision* criterion for diagnosis, the aspecific "tiredness" symptom may also be caused by physical illnesses. The Maier subscale does not include this item. Thus, the Maier subscale might especially be useful in patients having somatic complaints or illnesses. Additional methodological support comes from the exclusion of this somatic item by Santor and Coyne [21]. This hypothesis of better performance in patients with somatic complaints or illness needs further investigation, for

example, in comparison with the Hospital Anxiety and Depression Scale [63].

4.3. Previous comparative studies

Our findings are in line with findings of previous studies [11,20,26,40-44] and extend the evidence to support the Maier and Bech subscale as a valid alternative for the HDRS. This is relevant not only for the planning and conduction of clinical trials [40-42], but also for clinical practice [20,26]. Hooper and Bakish [26] found equal performance of the Bech subscale compared with the Montgomery Asberg Depression Rating Scale (MADRS) [30]. Because the MADRS was not used in our trials, we were unable to examine the performance of the Maier subscale compared with the MADRS. Hooper and Bakish [26] questioned whether a possible ceiling effect in the Bech subscale would limit its usefulness in severely depressed patients. In our study, more than 57% of the per protocol patients had an initial HDRS greater than 18 (indicative for severe depression) [11]. We did not find a ceiling effect in our diagnostic sample (data not shown) and found consequently higher E-S for the Maier and Bech subscales in initially severely depressed patients, indicative for an adequate sensitivity to measure (larger) changes caused by treatment. In addition to the observed ability to predict remission [41], we proposed cutoff scores for remission and various ranges for classification of depression severity.

In 2 publications, Bech et al [43,44] proposed the Bech subscale as an alternative measure to overcome the confounding influence of drug-related side effects in the comparison with placebo or active drugs. However, this problem is not fully solved, as tiredness may be induced by histaminergic effects from antidepressants (eg, tricyclics and mirtazapine) [43]. On the other hand, agitation (included in the Maier subscale) is known as an (mostly transient) SSRI-induced side effect.

An extra dimension of our study is that it extends the data for use of the Maier and Bech subscales in populations treated with psychotherapy. Hooper et al [26] and O'Sullivan et al [20] already demonstrated the usefulness of the Bech subscale in pharmacological treatment of melancholia, dysthymia, and typical and atypical depression.

An alarming point of our study is the moderate correlation of the Maier, Bech, and HDRS with the CGI-S and the CGI-I. Previous reports mentioned correlations between HDRS and CGI varying between 0.65 and 0.90 [11,28,40].Our results underscore the need of an HDRS or subscale rating instead of the CGI. We consider the validity of the CGI to be questionable, as most CGI raters (subjectively) evaluate their own treatment. Apparently, the clinician's judgement does not coincide with scale scores. In this respect, the performance of the (self-rated) SCL-90_{dep} is better. This was also illustrated in the ROC curves regarding the criterion of remission. Further research is needed to investigate whether correlations with the HDRS of other self-rated scales (eg, the Beck Depression Inventory [64]) are higher than the SCL-90_{dep}. In addition to this, a major limitation in our study and in any study investigating depression "severity" is that there is no definite gold standard. We used HDRS data as the gold standard, which means that scales under investigation can never be judged to be better than the HDRS; however this would be reversed if the CGI was used as a gold standard [65].

5. Conclusion

We think that both Maier and Bech subscales of the HDRS are equivalent to the HDRS and can easily be used to increase efficiency to measure treatment response in clinical practice. On theoretical grounds, we have a slight preference for the Maier subscale. The use of subscales would improve the efficiency and objectivity of measuring response in clinical practice, where often no scale (instead of a CGI) is used at all. This would further bridge the gap between clinical practice and research-based treatment recommendations for nonresponse in depression. Maier and Bech subscales should be compared in patients having comorbid somatic illnesses or patients treated with psychotherapy only. The impact of the difference of the one somatic item versus the agitation item between the Maier and Bech subscales and the consequences for their applicability in clinical subgroups needs further research.

Acknowledgment

The original randomized controlled trials were supported by an unrestricted educational grant from Eli Lilly Netherlands. All studies were carried out by the Mentrum Depression Research Group. The authors thank all psychotherapists, psychiatrists, and residents for their excellent work.

References

[1] Murray CJ, Lopez AD. Evidence-based health policy-lessons from the Global Burden of Disease Study. Science 1996;274:740-3.
[2] Greenberg PE, Stiglin LE, Finkelsteinc SN, Berndt ER. Depression: a neglected major illness. J Clin Psychiatry 1993;54:419-24.
[3] American Psychiatric Association. Practice guideline for the treatment of patients with major depressive disorder (revision). American Psychiatric Association. Am J Psychiatry 2000;157:1-45.
[4] Anderson IM, Nutt DJ, Deakin JF. Evidence-based guidelines for treating depressive disorders with antidepressants: a revision of the 1993 British Association for Psychopharmacology guidelines. J Psychopharmacol 2000;14:3-20.
[5] Mulrow CD, Williams Jr JW, Trivedi MH, Chiquette E, Aguilar C, Cornell JE, et al. Evidence report on: treatment of depression—newer pharmacotherapies. Psychopharmacol Bull 1999;34:409-795.
[6] Centraal Begeleidingsorgaan voor de Intercollegiale Toetsing. Consensusbijeenkomst depressie bij volwassenen. Utrecht: Centraal Begeleidingsorgaan voor de Intercollegiale Toetsing; 1994.
[7] Depression Guideline Panel. Depression in primary care: volume 1. Detection and diagnosis. Clinical practice guideline, number 5.

AHCPR publication no. 93-0550. Rockville (Md): U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research; 1993.
[8] Depression Guideline Panel. Depression in primary care: volume 2. Treatment of major depression. Clinical practice guideline, number 5. AHCPR publication no. 93-0551. Rockville (Md): U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research; 1993.
[9] Rush AJ, Crismon ML, Toprac MG, Trivedi MH, Rago WV, Shon S, et al. Consensus guidelines in the treatment of major depressive disorder. J Clin Psychiatry 1998;59:73-84.
[10] Trivedi MH, Rush AJ, Crismon ML, Kashner TM, Toprac MG, Carmody TJ, et al. Clinical results for patients with major depressive disorder in the Texas Medication Algorithm Project. Arch Gen Psychiatry 2004;61:669-80.
[11] Yonkers KA, Samson J. Mood disorders measures. In: Rush AJ, Pincus HA, First MB, et al, editors. Handbook of psychiatric measures, 1st ed. Washington (DC): American Psychiatric Association; 2000. p. 515-48.
[12] Ballenger JC. Clinical guidelines for establishing remission in patients with depression and anxiety. J Clin Psychiatry 1999; 60(Suppl 22):29-34.
[13] Biggs MM, Shores-Wilson K, Rush AJ, Carmody TJ, Trivedi MH, Crismon ML, et al. A comparison of alternative assessments of depressive symptom severity: a pilot study. Psychiatry Res 2000;96:269-79.
[14] Carroll BJ, Fielding JM, Blashki TG. Depression rating scales. A critical review. Arch Gen Psychiatry 1973;28:361-6.
[15] Anonymous. Scales for assessment of diagnosis and severity of mental disorders. Acta Psychiatr Scand Suppl 1993;372:1-87.
[16] Frank E, Prien RF, Jarrett RB, Keller MB, Kupfer DJ, Lavori PW, et al. Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Remission, recovery, relapse, and recurrence. Arch Gen Psychiatry 1991;48:851-5.
[17] Prien RF, Carpenter LL, Kupfer DJ. The definition and operational criteria for treatment outcome of major depressive disorder. A review of the current research literature. Arch Gen Psychiatry 1991; 48:796-800.
[18] Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23:56-61.
[19] Hamilton M. Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol 1967;6:278-96.
[20] O'Sullivan RL, Fava M, Agustin C, Baer L, Rosenbaum JF. Sensitivity of the six-item Hamilton Depression Rating Scale. Acta Psychiatr Scand 1997;95:379-84.
[21] Santor DA, Coyne JC. Examining symptom expression as a function of symptom severity: item performance on the Hamilton Rating Scale for Depression. Psychol Assess 2001;13:127-39.
[22] Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? Am J Psychiatry 2004;161:2163-77.
[23] Williams JB. Standardizing the Hamilton Depression Rating Scale: past, present, and future. Eur Arch Psychiatry Clin Neurosci 2001;251(Suppl 2):II6-12.
[24] Guy W, Clearly PA, Close JH, Conners CK, Covi L, et al. ECDEU Assessment manual for psychopharmacology. DHEW publication (ADM) 76-338. Washington (DC): US Department of Health, Education, and Welfare; 1976.
[25] American Psychiatric Association. Handbook of psychiatric measures. 1st ed. Washington (DC): American Psychiatric Association; 2000.
[26] Hooper CL, Bakish D. An examination of the sensitivity of the six-item Hamilton Rating Scale for Depression in a sample of patients suffering from major depressive disorder. J Psychiatry Neurosci 2000; 25:178-84.
[27] Linden M, Borchelt M, Barnow S, Geiselmann B. The impact of somatic morbidity on the Hamilton Depression Rating Scale in the very old. Acta Psychiatr Scand 1995;92:150-4.
[28] Maier W, Philipp M. Improving the assessment of severity of depressive states: a reduction of the Hamilton Depression Scale. Pharmacopsychiatry 1985;18:114-5.
[29] Gibbons RD, Clark DC, Kupfer DJ. Exactly what does the Hamilton Depression Rating Scale measure? J Psychiatr Res 1993; 27:259-73.
[30] Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry 1979;134:382-9.
[31] Maier W, Philipp M. Comparative analysis of observer depression scales. Acta Psychiatr Scand 1985;72:239-45.
[32] Moller HJ. Methodological aspects in the assessment of severity of depression by the Hamilton Depression Scale. Eur Arch Psychiatry Clin Neurosci 2001;251:13-20.
[33] Hughes JR, O'Hara MW, Rehm LP. Measurement of depression in clinical trials: an overview. J Clin Psychiatry 1982;43:85-8.
[34] Walczak DD, Apter JT, Halikas JA, Borison RL, Carman JS, Post GL, et al. The oral dose-effect relationship for fluvoxamine: a fixed-dose comparison against placebo in depressed outpatients. Ann Clin Psychiatry 1996;8:139-51.
[35] Moller HJ, Glaser K, Leverkus F, Gobel C. Double-blind, multicenter comparative study of sertraline versus amitriptyline in outpatients with major depression. Pharmacopsychiatry 2000;33:206-12.
[36] Bech P, Gram LF, Dein E, Jacobsen O, Vitger J, Bolwig TG. Quantitative rating of depressive states. Acta Psychiatr Scand 1975; 51:161-70.
[37] Bech P, Allerup P, Gram LF, Reisby N, Rosenberg R, Jacobsen O, et al. The Hamilton depression scale. Evaluation of objectivity using logistic models. Acta Psychiatr Scand 1981;63:290-9.
[38] Bech P, Allerup P, Reisby N, Gram LF. Assessment of symptom change from improvement curves on the Hamilton depression scale in trials with antidepressants. Psychopharmacology (Berl) 1984; 84:276-81.
[39] Bech P. The Bech-Rafaelsen Melancholia Scale (MES) in clinical trials of therapies in depressive disorders: a 20-year review of its use as outcome measure. Acta Psychiatr Scand 2002;106:252-64.
[40] Faries D, Herrera J, Rayamajhi J, DeBrota D, Demitrack M, Potter WZ. The responsiveness of the Hamilton Depression Rating Scale. J Psychiatr Res 2000;34:3-10.
[41] Entsuah R, Shaffer M, Zhang J. A critical examination of the sensitivity of unidimensional subscales derived from the Hamilton Depression Rating Scale to antidepressant drug effects. J Psychiatr Res 2002;36:437-48.
[42] Bech P, Tanghoj P, Andersen HF, Overo K. Citalopram dose-response revisited using an alternative psychometric approach to evaluate clinical effects of four fixed citalopram doses compared to placebo in patients with major depression. Psychopharmacology 2002;163:20-5.
[43] Bech P. Meta-analysis of placebo-controlled trials with mirtazapine using the core items of the Hamilton Depression Scale as evidence of a pure antidepressive effect in the short-term treatment of major depression. Int J Neuropsychopharmacol 2001;4:337-45.
[44] Bech P, Ciadella P, Haugh MC, Birkett MA, Hours A, Boissel JP, et al. Meta-analysis of randomised controlled trials of fluoxetine v. placebo and tricyclic antidepressants in the short-term treatment of major depression. Br J Psychiatry 2000;176:421-8.
[45] De Jonghe F, Kool S, Van Aalst G, Dekker J, Peen J. Combining psychotherapy and antidepressants in the treatment of depression. J Affect Disord 2001;64:217-29.
[46] Dekker J, Molenaar PJ, Kool S, Van Aalst G, Peen J, De Jonghe F. Dose-effect relations in time-limited combined psycho-pharmacological treatment for depression. Psychol Med 2005;35:47-58.
[47] Werman DS, editor. The practice of supportive psychotherapy. New York: Brunner/Mazel; 1984.
[48] Strupp HH, Binder JL. Psychotherapy in a new key. A guide to time-limited dynamic psychotherapy. New York: Basic Books; 1984.
[49] Rockland LH. Supportive therapy. A psychodynamic approach. New York: Basic Books; 1989.

[50] De Jonghe F, Rijnierse P, Janssen R. Psychoanalytic supportive psychotherapy. J Am Psychoanal Assoc 1994;42:421-46.
[51] Derogatis LR, Lipman RS, Covi L. SCL-90: an outpatient psychiatric rating scale—preliminary report. Psychopharmacol Bull 1973;9:13-28.
[52] Arindell WA, Ettema JM. Handleiding bij een multidimensionele psychopathologie-indicator. Lisse: Swets & Zeitlinger; 1986.
[53] de Jonghe F. Leidraad voor het scoren van de Hamilton Depression Rating Scale: HDRS leidraad. Amsterdam: Beneke Consultants; 1994.
[54] Fava M, Davidson KG. Definition and epidemiology of treatment-resistant depression. Psychiatr Clin North Am 1996;19:179-200.
[55] Drenth PJ, Sijtsma K. Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen. Houten: Bohn Stafleu Van Loghum; 1990.
[56] van Straten A, de Haan RJ, Limburg M, Schuling J, Bossuyt PM, van den Bos GA. A stroke-adapted 30-item version of the Sickness Impact Profile to assess quality of life (SA-SIP30). Stroke 1997;28:2155-61.
[57] Masters GN. A Rasch model for partial scoring. Psychometrika 1982;47:149-74.
[58] Verhelst ND, Glas CA, Verstralen HH. OPLM: computer program and manual. Arnhem (The Netherlands): CITO; 1995.
[59] Orlando M, Sherbourne CD, Thissen D. Summed-score linking using Item Response Theory: application to depression measurement. Psychol Assess 2000;12:354-9.
[60] Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003;54:573-83.
[61] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839-43.
[62] SPSS. SPSS for Windows [Release 10.1]. Chicago (Ill): SPSS Inc; 2000.
[63] Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983;67:361-70.
[64] Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561-71.
[65] Mulder RT, Joyce PR, Frampton C. Relationships among measures of treatment outcome in depressed patients. J Affect Disord 2003; 76:127-35.